PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 9/46 — A1

(11) International Publication Number: WO 99/44125

(43) International Publication Date: 2 September 1999 (02.09.99)

(21) International Application Number: PCT/US99/03394

(22) International Filing Date: 17 February 1999 (17.02.99)

(30) Priority Data:

60/076,048	26 February 1998 (26.02.98)	US
09/044,923	20 March 1998 (20.03.98)	US

(71) Applicant: SUN MICROSYSTEMS, INC. [US/US]; 901 San Antonio Road, MS UPAL01-521, Palo Alto, CA 94303 (US).

(72) Inventors: WOLLRATH, Ann, M.; 9 Northwoods Road, Groton, MA 01450 (US). WALDO, James, H.; 155 Ruby Road, Dracut, MA 01826 (US). ARNOLD, Kenneth, C., R., C.; 7 Moon Hill Road, Lexington, MA 02173 (US).

(74) Agents: GARRETT, Arthur, S.; Finnegan, Henderson, Farrabow, Garrett & Dunner, L.L.P., 1300 I Street, N.W., Washington, DC 20005-3315 (US) et al.

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD AND SYSTEM FOR LEASING STORAGE

(57) Abstract

A method and system for leasing storage locations in a distributed processing system is provided. Consistent with this method and system, a client requests access to storage locations for a period of time (lease period) from a server, such as the file system manager. Responsive to this request, the server invokes a lease period algorithm, which considers various factors to determine a lease period during which time the client may access the storage locations. After a lease is granted, the server sends an object to the client that advises the client of the lease period and that provides the client with behavior to modify the lease, like canceling the lease or renewing the lease. The server supports concurrent leases, exact leases, and leases for various types of access. After all leases to a storage location expire, the server reclaims the storage location.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

METHOD AND SYSTEM FOR LEASING STORAGE

RELATED APPLICATIONS

This is a continuation-in-part of U.S. Patent Application No. 08/729,421, filed on October 11, 1996, which is incorporated herein by reference.

The following identified U.S. patent applications are relied upon and are incorporated by reference in this application.

Provisional U.S. Patent Application No. 60/076,048, entitled "Distributed Computing System," filed on February 26, 1998.

U.S. Patent Application No. 09/044,838, entitled "Method, Apparatus, and Product for Leasing of Delegation Certificates in a Distributed System," bearing attorney docket no. 06502.0011-02000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,834, entitled "Method, Apparatus and Product for Leasing of Group Membership in a Distributed System," bearing attorney docket no. 06502.0011-03000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,916, entitled "Leasing for Failure Detection," bearing attorney docket no. 06502.0011-04000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,933, entitled "Method for Transporting Behavior in Event Based System," bearing attorney docket no. 06502.0054-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,919, entitled "Deferred Reconstruction of Objects and Remote Loading for Event Notification in a Distributed System," bearing attorney docket no. 06502.0062-01000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,938, entitled "Methods and Apparatus for Remote Method Invocation," bearing attorney docket no. 06502.0102-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/045,652, entitled "Method and System for Deterministic Hashes to Identify Remote Methods," bearing attorney docket no. 06502.0103-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,790, entitled "Method and Apparatus for Determining Status of Remote Objects in a Distributed System," bearing attorney docket no. 06502.0104-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,930, entitled "Downloadable Smart Proxies for Performing Processing Associated with a Remote Procedure Call in a Distributed System," bearing attorney docket no. 06502.0105-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,917, entitled "Suspension and Continuation of Remote Methods," bearing attorney docket no. 06502.0106-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,835, entitled "Method and System for Multi-Entry and Multi-Template Matching in a Database," bearing attorney docket no. 06502.0107-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,839, entitled "Method and System for In-Place Modifications in a Database," bearing attorney docket no. 06502.0108, and filed on the same date herewith.

U.S. Patent Application No. 09/044,945, entitled "Method and System for Typesafe Attribute Matching in a Database," bearing attorney docket no. 06502.0109-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,931, entitled "Dynamic Lookup Service in a Distributed System," bearing attorney docket no. 06502.0110-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,939, entitled "Apparatus and Method for Providing Downloadable Code for Use in Communicating with a Device in a Distributed System," bearing attorney docket no. 06502.0112-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,826, entitled "Method and System for Facilitating Access to a Lookup Service," bearing attorney docket no. 06502.0113-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,932, entitled "Apparatus and Method for Dynamically Verifying Information in a Distributed System," bearing attorney docket no. 06502.0114-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/030,840, entitled "Method and Apparatus for Dynamic Distributed Computing Over a Network," and filed on February 26, 1998.

U.S. Patent Application No. 09/044,936, entitled "An Interactive Design Tool for Persistent Shared Memory Spaces," bearing attorney docket no. 06502.0116-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,934, entitled "Polymorphic Token-Based Control," bearing attorney docket no. 06502.0117-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044.915, entitled "Stack-Based Access Control," bearing attorney docket no. 06502.0118-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,944, entitled "Stack-Based Security Requirements," bearing attorney docket no. 06502.0119-00000, and filed on the same date herewith.

U.S. Patent Application No. 09/044,837, entitled "Per-Method Designation of Security Requirements," bearing attorney docket no. 06502.0120-00000, and filed on the same date herewith.

BACKGROUND OF THE INVENTION

A. Field of the Invention

This invention generally relates to data processing systems and, more particularly, to leasing storage in data processing systems.

B. Description of the Related Art

Proper resource management is an important aspect to efficient and effective use of computers. In general, resource management involves allocating resources (e.g., memory) in response to requests as well as deallocating resources at appropriate times, for example, when the requesters no longer require the resources. In general, the resources contain data referenced by computational entities (e.g., applications, programs, applets, etc.) executing in the computers.

In practice, when applications executing on computers seek to refer to resources, the computers must first allocate or designate resources so that the applications can properly refer to them. When the applications no longer refer to a resource, the computers can deallocate or reclaim the resource for reuse. In computers each resource has a unique "handle" by which the resource can be referenced. The handle may be implemented in various ways, such as an address, array index, unique value, pointer, etc.

Resource management is relatively simple for a single computer because the events indicating when resources can be reclaimed, such as when applications no longer refer to them or after a power failure, are easy to determine. Resource management for distributed systems connecting multiple computers is more difficult because applications in several different computers may be using the same resource.

Disconnects in distributed systems can lead to the improper and premature reclamation of resources or to the failure to reclaim resources. For example, multiple applications operating on different computers in a distributed system may refer to resources located on other machines. If connections between the computers on which resources are located and the applications referring to those resources are interrupted, then the computers may reclaim the resources prematurely. Alternatively, the computers may maintain the resources in perpetuity, despite the extended period of time that applications failed to access the resources.

These difficulties have led to the development of systems to manage network resources, one of which is known as "distributed garbage collection." That term describes a facility provided by a language or runtime system for distributed systems that automatically manages resources used by an application or group of applications running on different computers in a network.

In general, garbage collection uses the notion that resources can be freed for future use when they are no longer referenced by any part of an application. Distributed garbage collection extends this notion to the realm of distributed computing, reclaiming resources when no application on any computer refers to them.

Distributed garbage collection must maintain integrity between allocated resources and the references to those resources. In other words, the system must not be permitted to deallocate or free a resource when an application running on any computer in the network continues to refer to that resource. This reference-to-resource binding, referred to as "referential integrity," does not guarantee that the reference will always grant access to the resource to which it refers. For example, network failures can make such access impossible. The integrity, however, guarantees that if the reference can be used to gain access to any resource, it will be the same resource to which the reference was first given.

Distributed systems using garbage collection must also reclaim resources no longer being referenced at some time in the finite future. In other words, the system must provide a guarantee

against "memory leaks." A memory leak can occur when all applications drop references to a resource, but the system fails to reclaim the resource for reuse because, for example, of an incorrect determination that some application still refers to the resource.

Referential integrity failures and memory leaks often result from disconnections between applications referencing the resources and the garbage collection system managing the allocation and deallocation of those resources. For example, a disconnection in a network connection between an application referring to a resource and a garbage collection system managing that resource may prevent the garbage collection system from determining whether and when to reclaim the resource. Alternatively, the garbage collection system might mistakenly determine that, since an application has not accessed a resource within a predetermined time, it may collect that resource. A number of techniques have been used to improve the distributed garbage collection mechanism by attempting to ensure that such mechanisms maintain referential integrity without memory leaks. One conventional approach uses a form of reference counting, in which a count is maintained of the number of applications referring to each resource. When a resource's count goes to zero, the garbage collection system may reclaim the resource. Such a reference counting scheme only works, however, if the resource is created with a corresponding reference counter. The garbage collection system in this case increments the resource's reference count as additional applications refer to the resource, and decrements the count when an application no longer refers to the resource.

Reference counting schemes, however, especially encounter problems in the face of failures that can occur in distributed systems. Such failures can take the form of a computer or application failure or network failure that prevent the delivery of messages notifying the garbage collection system that a resource is no longer being referenced. If messages go undelivered because of a network disconnect, the garbage collection system does not know when to reclaim the resource.

To prevent such failures, some conventional reference counting schemes include "keep-alive" messages, which are also referred to as "ping back." According to this scheme, applications in the network send messages to the garbage collection system overseeing resources and indicate that the applications can still communicate. These messages prevent the garbage collection system from dropping references to resources. Failure to receive such a "keep-alive" message indicates that the garbage collection system can decrement the reference count for a

resource and, thus, when the count reaches zero, the garbage collection system may reclaim the resource. This, however, can still result in the premature reclamation of resources following reference counts reaching zero from a failure to receive "keep-alive" messages because of network failures. This violates the referential integrity requirement.

Another proposed method for resolving referential integrity problems in garbage collection systems is to maintain not only a reference count but also an identifier corresponding to each computational entity referring to a resource. See A. Birrell, et al., "Distributed Garbage Collection for Network Objects," No. 116, digital Systems Research Center, December 15, 1993. This method suffers from the same problems as the reference counting schemes. Further, this method requires the addition of unique identifiers for each computational entity referring to each resource, adding overhead that would unnecessarily increase communication within distributed systems and add storage requirements (i.e., the list of identifiers corresponding to applications referring to each resource).

SUMMARY OF THE INVENTION

In accordance with the present invention, referential integrity is guaranteed without costly memory leaks by leasing resources for a period of time during which the parties in a distributed system, for example, an application holding a reference to a resource and the garbage collection system managing that resource, agree that the resource and a reference to that resource will be guaranteed. At the end of the lease period, the guarantee that the reference to the resource will continue lapses, allowing the garbage collection system to reclaim the resource. Because the application holding the reference to the resource and the garbage collection system managing the resource agree to a finite guaranteed lease period, both can know when the lease and, therefore, the guarantee, expires. This guarantees referential integrity for the duration of a reference lease and avoids the concern of failing to free the resource because of network errors. In addition to memory, the leasing technique can be applied to other types of storage, such as storage devices.

Consistent with an alternative embodiment of the present invention, as embodied and broadly described herein, a method for leasing storage locations is provided. This method comprises the steps of receiving a request from a caller specifying a storage location and a lease period, determining a lease period during which the caller has access to the specified storage

locations, advising the caller of the granted lease period, and permitting the caller to access storage locations for the determined lease period.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an embodiment of the invention and, together with the description, serve to explain the advantages and principles of the invention. In the drawings,

FIG. 1 is a flow diagram of the steps performed by the application call processor according to an implementation of the present invention;

FIG. 2 is a flow diagram of the steps performed by the server call processor to process dirty calls according to the implementation of the present invention;

FIG. 3 is a flow diagram of the steps performed by the server call processor to process clean calls according to the implementation of the present invention;

FIG. 4 is a flow diagram of the steps performed by the server call processor to initiate a garbage collection process according to the implementation of the present invention.

FIG. 5 is a diagram of a preferred flow of calls within a distributed processing system;

FIG. 6 is a block diagram of the components of the implementation of a method invocation service according to the present invention;

FIG. 7 is a diagram of a distributed processing system that can be used in an implementation of the present invention; and

FIG. 8 is a diagram of the individual software components in the platforms of the distributed processing system according to the implementation of the present invention; and

FIG. 9 is a diagram of a data processing system for leasing storage locations in a distributed processing system that can be used in an alternative embodiment of the present invention; and

FIG. 10A and FIG. 10B represent a flow diagram of the steps performed by a client when requesting a lease from a server according to an alternative embodiment of the present invention; and

FIG. 11 is a flow diagram of the steps performed by a server when a client requests a lease according to an alternative embodiment of the present invention.

DETAILED DESCRIPTION

Reference will now be made in detail to an implementation of the present invention as illustrated in the accompanying drawings. Wherever possible, the same reference numbers will be used throughout the drawings and the following description to refer to the same or like parts.

The present invention may be implemented by computers organized in a conventional distributed processing system architecture. The architecture for and procedures to implement this invention, however, are not conventional, because they provide a distributed garbage collection cheme that ensures referential integrity and eliminates memory leaks.

A. Overview

A method invocation (MI) component located in each of the computers in the distributed processing system implements the distributed garbage collection scheme of this invention. The MI component may consist of a number of software modules preferably written in the JAVA™ programming language.

In general, whenever an application in the distributed processing system obtains a reference to a distributed resource, by a name lookup, as a return value to some other call, or another method, and seeks to access the resource, the application makes a call to the resource or to an MI component managing the resource. That MI component, called a managing MI component, keeps track of the number of outstanding references to the resource. When the number of references to a reference is zero, the managing MI component can reclaim the resource. The count of the number of references to a resource is generally called the "reference count" and the call that increments the reference count may be referred to as a "dirty call."

When an application no longer requires a distributed resource, it sends a different call to the resource or the managing MI component. Upon receipt of this call, the managing MI component decrements the reference count for the resource. This call to drop a reference may be referred to as a "clean call."

In accordance with an implementation of the present invention, a dirty call can include a requested time interval, called a lease period, for the reference to the resource. Upon receipt of the dirty call, the managing MI component sends a return call indicating a period for which the lease was granted. The managing MI component thus tracks the lease period for those references as well as the number of outstanding references. Consequently, when the reference

count for a resource goes to zero or when the lease period for the resource expires, the managing MI component can reclaim the resource.

B. Procedure

An application call processor in an MI component performs the steps of the application call procedure 100 illustrated in FIG. 1. The server call processor in the managing MI component performs the steps of the procedures 200, 300, and 400 illustrated in FIGs. 2-4, respectively. The managing MI component's garbage collector performs conventional procedures to reclaim resources previously bound to references in accordance with instructions from the server call processor. Accordingly, the conventional procedures of the garbage collector will not be explained.

1. Application Call Processor

FIG. 1 is a flow diagram of the procedure 100 that the application call processor of the MI component uses to handle application requests for references to resources managed by the same or another MI component located in the distributed processing system.

After an application has obtained a reference to a resource, the application call processor sends a dirty call, including the resource's reference and a requested lease period to the managing MI component for the resource (step 110). The dirty call may be directed to the resource itself or to the managing MI component.

The application call processor then waits for and receives a return call from the managing MI component (step 120). The return call includes a granted lease period during which the managing MI component guarantees that the reference of the dirty call will be bound to its resource. In other words, the managing MI component agrees not to collect the resource corresponding to the reference of a dirty call for the grant period. If the managing MI component does not provide a grant period, or rejects the request for a lease, then the application call processor will have to send another dirty call until it receives a grant period.

The application call processor monitors the application's use of the reference and, either when the application explicitly informs the application call processor that the reference is no longer required or when the application call processor makes this determination on its own (step 130), the application call processor sends a clean call to the managing MI component (step 140).

In a manner similar to the method used for dirty calls, the clean call may be directed to the referenced resource and the managing MI component will process the clean call. Subsequently, the application call processor eliminates the reference from a list of references being used by the application (step 150).

If the application is not yet done with the reference (step 130), but the application call processor determines that the grant period for the reference is about to expire (step 160), then the application call processor repeats steps 110 and 120 to ensure that the reference to the resource is maintained by the managing MI component on behalf of the application.

2. Server Call Processor

The MI component's server call processor performs three main procedures: (1) handling dirty calls; (2) handling incoming clean calls; and (3) initiating a garbage collection cycle to reclaim resources at the appropriate time.

(I) Dirty Calls

FIG. 2 is a flow diagram of the procedure 200 that the MI component's server call processor uses to handle requests to reference resources, i.e., dirty calls, that the MI software component manages. These requests come from application call processors of MI components in the distributed processing system, including the application call processor of the same MI component as the server call processor handling requests.

First, the server call processor receives a dirty call (step 210). The server call processor then determines an acceptable grant period (step 220). The grant period may be the same as the requested lease period or some other time period. The server call processor determines the appropriate grant period based on a number of conditions including the amount of resource required and the number of other grant periods previously granted for the same resource.

When the server call processor determines that a resource has not yet been allocated for the reference of a dirty call (step 230), the server call processor allocates the required resource (step 240).

The server call processor then increments a reference count corresponding to the reference of a dirty call (step 250), sets the acceptable grant period for the reference-to-resource binding (step 260), and sends a return call to an application call processor with the grant period

(step 270). In this way, the server call processor controls incoming dirty calls regarding references to resources under its control.

Applications can extend leases by sending dirty calls with an extension request before current leases expire. As shown in procedure 200, a request to extend a lease is treated just like an initial request for a lease. An extension simply means that the resource will not be reclaimed for some additional interval of time, unless the reference count goes to zero.

(ii) Clean Calls

The MI component's server call processor also handles incoming clean calls from application call processors. When an application in the distributed processing system no longer requires a reference to a resource, it informs the MI component managing the resource for that reference so that the resource may be reclaimed for reuse. Fig. 3 is a flow diagram of the procedure 300 with the steps that the MI component's server call processor uses to handle clean calls.

When the server call processor receives a clean call with a reference to a resource that the MI component manages (step 310), the server call processor decrements a corresponding reference count (step 320). The clean call may be sent to the resource, with the server call processor monitoring the resource and executing the procedure 300 to process the call. Subsequently, the server call processor sends a return call to the MI component that sent the clean call to acknowledge receipt (step 330). In accordance with this implementation of the present invention, a clean call to drop a reference may not be refused, but it must be acknowledged.

(iii) Garbage Collection

The server call processor also initiates a garbage collection cycle to reclaim resources for which it determines that either no more references are being made to the resource or that the agreed lease period for the resource has expired. The procedure 400 shown in FIG. 4 includes a flow diagram of the steps that the server call processor uses to initiate a garbage collection cycle.

The server call processor monitors reference counts and granted lease periods and determines whether a reference count is zero for a resource managed by the MI component, or

the grant period for a reference has expired (step 410). When either condition exists, the server call processor initiates garbage collection (step 420) of that resource. Otherwise, the server call processor continues monitoring the reference counts and granted lease periods.

C. Call Flow

FIG. 5 is a diagram illustrating the flow of calls among MI components within the distributed processing system. Managing MI component 525 manages the resources 530 by monitoring the references to those resources 530 (see garbage collect 505). Because the managing MI components 525 manages the resources, the server call processor of managing MI component 525 performs the operations of this call flow description.

FIG. 5 also shows that applications 510 and 540 have corresponding MI components 515 and 545, respectively. Each of the applications 510 and 540 obtains a reference to one of the resources 530 and seeks to obtain access to one of the resources 530 such that a reference is bound to the corresponding resource. To obtain access, applications 510 and 540 invoke their corresponding MI components 515 and 545, respectively, to send dirty calls 551 and 571, respectively, to the MI component 525. Because the MI components 515 and 525 handle application requests for access to resources 530 managed by another MI component, such as managing MI component 525, the application call processors of MI components 515 and 545 perform the operations of this call flow description.

In response to the dirty calls 551 and 571, managing MI component 525 sends return calls 552 and 572, respectively, to each of the MI components 515 and 545, respectively. The dirty calls include granted lease periods for the references of the dirty calls 551 and 571.

Similarly, FIG. 5 also shows MI components 515 and 545 sending clean calls 561 and 581, respectively, to managing MI component 525. Clean calls 561 and 581 inform managing MI component 525 that applications 510 and 540, respectively, no longer require access to the resource specified in the clean calls 561 and 581. Managing MI component 525 responds to clean calls 561 and 581 with return calls 562 and 582, respectively. Return calls 562 and 582 differ from return calls 552 and 572 in that return calls 562 and 582 are simply acknowledgments from MI component 525 of the received clean calls 561 and 581.

Both applications 510 and 540 may request access to the same resource. For example, application 510 may request access to "RESOURCE(1)" while application 540 was previously

granted access to that resource. MI component 525 handles this situation by making the resource available to both applications 510 and 540 for agreed lease periods. Thus, MI component 525 will not initiate a garbage collection cycle to reclaim the "RESOURCE(1)" until either applications 510 and 540 have both dropped their references to that resource or the latest agreed periods has expired, whichever event occurs first.

By permitting more than one application to access the same resource simultaneously, the present invention also permits an application to access a resource after it sent a clean call to the managing MI component dropping the reference to the resource. This occurs because the resource is still referenced by another application or the reference's lease has not yet expired so the managing MI component 525 has not yet reclaimed the resource. The resource, however, will be reclaimed after a finite period, either when no more applications have leases or when the last lease expires.

D. MI Components

FIG. 6 is a block diagram of the modules of an MI component 600 according to an implementation of the present invention. MI component 600 can include a reference component 605 for each reference monitored, application call processor 640, server call processor 650, and garbage collector 660.

Reference component 605 preferably constitutes a table or comparable structure with reference data portions 610, reference count 620, and grant period register 630. MI component 600 uses the reference count 620 and grant period 630 for each reference specified in a corresponding reference data portion 610 to determine when to initiate garbage collector 660 to reclaim the corresponding resource.

Application call processor 640 is the software module that performs the steps of procedure 100 in FIG. 1. Server call processor 650 is the software module that performs the steps of procedures 200, 300, and 400 in FIGs. 2-4. Garbage collector 660 is the software module that reclaims resources in response to instructions from the server call processor 650, as explained above.

E. Distributed Processing System

FIG. 7 illustrates a distributed processing system 50 which can be used to implement the present invention. In FIG. 7, distributed processing system 50 contains three independent and heterogeneous platforms 100, 200, and 300 connected in a network configuration represented by the network cloud 55. The composition and protocol of the network configuration represented in FIG. 7 by the cloud 55 is not important as long as it allows for communication of the information between platforms 700, 800 and 900. In addition, the use of just three platforms is merely for illustration and does not limit the present invention to the use of a particular number of platforms. Further, the specific network architecture is not crucial to this invention. For example, another network architecture that could be used in accordance with this invention would employ one platform as a network controller to which all the other platforms would be connected.

In the implementation of distributed processing system 50, platforms 700, 800 and 900 each include a processor 710, 810, and 910 respectively, and a memory, 750, 850, and 950, respectively. Included within each processor 710, 810, and 910, are applications 720, 820, and 920, respectively, operating systems 740, 840, and 940, respectively, and MI components 730, 830, and 930, respectively.

Applications 720, 820, and 920 can be programs that are either previously written and modified to work with the present invention, or that are specially written to take advantage of the services offered by the present invention. Applications 720, 820, and 920 invoke operations to be performed in accordance with this invention.

MI components 730, 830, and 930 correspond to the MI component 600 discussed above with reference to FIG. 6.

Operating systems 740, 840, and 940 are standard operating systems tied to the corresponding processors 710, 810, and 910, respectively. The platforms 700, 800, and 900 can be heterogenous. For example, platform 700 has an UltraSparc® microprocessor manufactured by Sun Microsystems Corp. as processor 710 and uses a Solaris® operating system 740. Platform 800 has a MIPS microprocessor manufactured by Silicon Graphics Corp. as processor 810 and uses a Unix operating system 840. Finally, platform 900 has a Pentium microprocessor manufactured by Intel Corp. as processor 910 and uses a Microsoft Windows 95 operating system

940. The present invention is not so limited and could accommodate homogenous platforms as well.

Sun, Sun Microsystems, Solaris, Java, and the Sun Logo are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and other countries. UltraSparc and all other SPARC trademarks are used under license and are trademarks of SPARC International, Inc. in the United States and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc.

Memories 750, 850, and 950 serve several functions, such as general storage for the associated platform. Another function is to store applications 720, 820, and 920, MI components 730, 830, and 930, and operating systems 740, 840, and 940 before execution by the respective processor 710, 810, and 910. In addition, portions of memories 750, 850, and 950 may constitute shared memory available to all of the platforms 700, 800, and 900 in network 50.

E. MI Services

The present invention may be implemented using a client/server model. The client generates requests, such as the dirty calls and clean calls, and the server responds to requests.

Each of the MI components 730, 830 and 930 shown in FIG. 7 preferably includes both client components and server components. FIG. 8, which is a block diagram of a client platform 1000 and a server platform 1100, applies to any two of the platforms 700, 800, and 900 in FIG. 7.

Platforms 1000 and 1100 contain memories 1050 and 1150, respectively, and processors 1010 and 1110, respectively. The elements in the platforms 1000 and 1100 function in the same manner as similar elements described above with reference to FIG. 7. In this example, processor 1010 executes a client application 1020 and processor 1110 executes a server application 1120. Processors 1010 and 1110 also execute operating systems 1040 and 1140, respectively, and MI components 1030 and 1130, respectively.

MI components 1030 and 1130 each include a server call processor 1031 and 1131, respectively, an application call processor 1032 and 1132, respectively, and a garbage collector 1033 and 1133, respectively. Each of the MI components 1030 and 1130 also contains reference components, including reference data portions 1034 and 1134, respectively, reference counts

1035 and 1135, respectively, and grant period registers 1036 and 1136, respectively, for each reference that the respective MI component 1030 or 1130 monitors.

Application call processors 1032 and 1132 represent the client service and communicate with server call processors 1031 and 1131, respectively, which represent the server service. Because platforms 1000 and 1100 contain a server call processor, an application call processor, a garbage collector, and reference components, either platform can act as a client or a server.

For purposes of the discussion that follows, however, platform 1000 is designated the client platform and platform 1100 is designated as the server platform. In this example, client application 1020 obtains references to distributed resources and uses MI component 1030 to send dirty calls to the resources managed by MI component 1130 of server platform 1100.

Additionally, server platform 1100 may be executing a server application 1120. Server application 1120 may also use MI component 1130 to send dirty calls, which may be handled by MI component 1130 when the resources of those dirty calls are managed by MI component 1130. Alternatively, server application 1120 may use MI component 1130 to send dirty calls to resources managed by MI component 1030.

Accordingly, server call processor 1031, garbage collector 1033, and reference count 1035 for MI component 1030 of client platform 1000 are not active and are therefore presented in FIG. 8 as shaded. Likewise, application call processor 1132 of MI component 1130 of the server platform 1100 is shaded because it is also dormant.

When client application 1020 obtains a reference corresponding to a resource, application call processor 1032 sends a dirty call, which server call processor 1131 receives. The dirty call includes a requested lease period. Server call processor 1131 increments the reference count 1135 for the reference in the dirty call and determines a grant period. In response, server call processor 1131 sends a return call to application call processor 1030 with the grant period. Application call processor 1032 uses the grant period to update recorded grant period 1035, and to determine when the resource corresponding to the reference of its dirty call may be reclaimed.

Server call processor 1131 also monitors the reference counts and grant periods corresponding to references for resources that it manages. When one of its reference counts 1135 is zero, or when the grant period 1135 for a reference has expired, whichever event occurs first, server call processor 1131 may initiate the garbage collector 1133 to reclaim the resource corresponding to the reference that has a reference count of zero or an expired grant period.

The leased-reference scheme according to the implementation of the present invention does not require that the clocks on the platforms 1000 and 1100 involved in the protocol be synchronized. The scheme merely requires that they have comparable periods of increase. Leases do not expire at a particular time, but rather expire after a specific time interval. As long as there is approximate agreement on the interval, platforms 1000 and 1100 will have approximate agreement on the granted lease period. Further, since the timing for the lease is, in computer terms, fairly long, minor differences in clock rate will have little or no effect.

The transmission time of the dirty call can affect the protocol. If MI component 1030 holds a lease to reference and waits until just before the lease expires to request a renewal, the lease may expire before the MI component 1130 receives the request. If so, MI component 1130 may reclaim the resource before receiving the renewal request. Thus, when sending dirty calls, the sender should add a time factor to the requested lease period in consideration of transmission time to the platform handling the resource of a dirty call so that renewal dirty calls may be made before the lease period for the resource expires.

F. Conclusion

In accordance with the present invention a distributed garbage collection scheme ensures referential integrity and eliminates memory leaks by providing granted lease periods corresponding to references to resources in the distributed processing system such that when the granted lease periods expire, so do the references to the resources. The resources may then be collected. Resources may also be collected when they are no longer being referenced by processes in the distributed processing system with reference to counters assigned to the references for the resources.

Alternative Embodiment of the Present Invention

The leasing technique, described above, relates to garbage collection. However, an alternative embodiment of the present invention, as described below, can be used with storage devices.

Storage devices have many storage locations containing various logical groupings of data that may be used by more than one program. These logical groupings may take the form of files, databases, or documents. The leasing of storage locations allows access (e.g., read and write

access) to the storage locations for a pre-negotiated amount of time. It is immaterial to the leasing of storage locations what kind of data is contained in the storage locations or whether the storage locations contain any data at all. Also, the leasing of storage locations can be applied on different levels of storage, such as database fields, files, blocks of storage, or actual storage locations.

In a computer system or a distributed system, many programs may compete for files stored in various groups of storage locations. Thus, groups of storage locations may have many programs vying for access. The leasing technique can be used to arbitrate the use of storage locations in such an environment.

When using a lease for a group of storage locations containing the data for a file, a program ("the client") requests a lease from the file system manager ("the server") to access the group of storage locations for a period of time ("the lease period"). Depending on availability, priority, and other factors described below, the server either denies the request or grants a lease period. The lease period granted may be either the entire lease period requested or some portion of it. Once a client receives a lease, the client may access the group of storage locations for the lease period.

When requesting a lease period, the client may request an exact lease period. In this situation, the server only grants the lease if the lease period would be the entire lease period requested, as opposed to a portion of it.

While a lease is active, the client is guaranteed access to the group of storage locations and may perform read and write operations on them. And, likewise, the server, during an active lease, will maintain the storage locations' integrity. For example, during the lease period, the server will not allow the leased file to be deleted, written over, or otherwise affected by any entity other than the client. After a lease expires, however, the server no longer guarantees the integrity of the file to the client, and thus, the server may delete the file or otherwise materially change it, or grant a lease to another client that may do the same. Storage locations with no outstanding leases are reclaimed by the server.

Each storage location may have an associated limiting parameter, such as an access parameter or a privilege parameter. The access parameter determines the type of access the server supports for that storage location. For example, a storage location may be defined as read-access only. In this case, the server only allows read access for a subsequently granted lease for

that particular storage location. Conversely, an attempt by the client to write to that storage location would not be permitted by the server. Other potential storage location access parameters may include write access, allocation access, re-allocation access, and sub-block access (i.e., for large blocks of storage).

The associated privilege parameter specifies the privilege level the client must have before a lease will be granted. The server may use the privilege parameter to prioritize competing lease requests. In other words, when the server has multiple outstanding lease requests for the same storage location, it may prioritize the requests based on the privilege level of the clients making the request.

The alternative embodiment also supports concurrent access to a group of storage locations by granting multiple, concurrent leases to the same storage location. For example, if a particular storage location's parameter specifies "read" access, the server can grant multiple concurrent leases to that storage location without breaching the integrity of the storage location. Concurrent leases could also be applied, for example, to large files. The server could merely grant leases to smaller sub-blocks of the file, again, without compromising the integrity of the larger file.

Once the client requests a lease, the server returns to the client an object, including methods for determining the duration of the lease, for renewing the lease, and for canceling the lease. The object is an instance of a class that may be extended in many ways to offer more functionality, but the basic class is defined as follows:

```
interface Lease {
        obj FileHandle;
        public long getDuration ( );
        public void cancel ( ) throws  UnknownLeaseException,
                                       RemoteException;
        public void renew (long renewDuration) throws
                                LeaseDeniedException,
                                UnknownLeaseException,
                                RemoteException;
        }
```

Specifically, invoking the "getDuration" method provides the client with the length of the granted lease period. This period represents the most recent lease granted by the server. It is the client's responsibility, however, to determine the amount of time outstanding on the lease.

The "renew" method permits the client to renew the lease, asking for more time, without having to re-initiate the original lease request. Situations where the client may desire to renew the lease include when the original lease proves to be insufficient (i.e., the client requires additional use of the storage location), or when only a partial lease (i.e., less than the requested lease) was granted.

The client may use the renew method to request an additional lease period, or the client may continually invoke the renew method multiple times until many additional lease periods are granted. The renew method has no return value; if the renewal is granted, the new lease period will be reflected in the lease object on which the call was made. If the server is unable or unwilling to renew the lease, the reason is set forth in the LeaseDeniedException generated by the renew method.

Finally, the client invokes the "cancel" method when the client wishes to cancel the lease, but there is still time left on the lease. Thus, cancel allows the server to re-claim the storage locations so that other programs make access them. Accordingly, the cancel method ensures that the server can optimize the use of the storage locations in the distributed system. In contrast, upon the end of a lease (i.e., natural termination), the server knows to take back control of storage locations. Therefore, the client has no obligation to notify the server upon the natural termination of the lease.

Figure 9 depicts a data processing system 9000 suitable for use by an alternative embodiment of the present invention. The data processing system 9000 includes a computer system 9001 connected to the Internet 9002. The computer system 9001 includes a memory 9003, a secondary storage device 9004, a central processing unit (CPU) 9006, an input device 9008, and a video display 9010. The memory 9003 further includes an operating system 9012 and a program ("the client") 9014. The operating system 9012 contains a file system manager ("the server") 9016 that manages files 9018 on the secondary storage device 9004. The client 9014 requests access to one or more of the files 9018 by requesting a lease from the server 9016. In response, the server 9016 may either choose to grant or deny the lease as further described below. One skilled in the art will appreciate that computer 9000 may contain additional or different components.

Although aspects of the alternative embodiment are described as being stored in memory 9003, one skilled in the art will appreciate that these aspects may also be stored in other

computer-readable media, such as secondary storage devices, like hard disks, floppy disks, or CD-Rom; a carrier wave from the Internet 9002; or other forms of RAM or ROM. Additionally, one skilled in the art will appreciate that the alternative embodiment can be used to lease other forms of data in secondary storage, like databases, spreadsheets, and documents.

Figures 10A and 10B depict a flowchart of the steps performed by the client when requesting a lease from the server. The first step performed by the client is to send a request for a lease to the server (step 10002). This request is a function call that includes a number of parameters, including (1) the requested storage locations the client wishes to lease, (2) the desired lease period, (3) an exact lease indicator, (4) the type of access the client desires, and (5) the client's privilege.

The requested storage location contains an indication of the storage locations to be leased. The desired lease period contains an amount of time that the client wants to utilize the storage locations. The exact lease request contains an indication of whether an exact lease request is being made or whether a lease of less than the requested amount will suffice. The type of access requested indicates the type of storage location access the client requested. The types of access include read access, write access, allocation access, re-allocation access, and sub-block access (i.e., for large blocks of storage). The privilege field indicates the privilege level of the user or the client. To form a valid request, the client request must contain both the requested storage location and the desired lease period.

There are two general scenarios that generate a lease request for storage locations. The first scenario occurs when a file is created. The creation command used to create the file also generates a lease request to the server for access to the file. The requirement that a new file be regulated by the lease technique ensures that storage locations do not remain unaccounted for. Thus, the server would not be inclined to grant long or indefinite leases for new files. The second scenario occurs when a client desires access to existing storage locations or a file already having an existing lease (i.e., in the case of concurrent leases).

After sending the request, the client determines if a lease was granted by whether it receives a lease object from the server (step 10006). The lease object contains, various information, as described above, including the file handle, the getDuration method, the renew method, and the cancel method. It should be noted that if the server rejects the lease for any

reason, the server generates an exception, which is handled by various exception handlers of the client.

If a lease was not granted because of an improper request (step 10008), an exception handler of the client is invoked that reconfigures the request (step 10010) and processing continues to step 10002. An improper request includes lease period requests that are too large or lease requests for an unknown storage location. If the request was improper, the client reconfigures the request to create a valid request. For example, if the server was unable to grant an exact lease request, the client may reconfigure the request to reflect a lesser lease period, or if the request was for an unknown storage location, the client may reconfigure the request to reflect a known storage location.

However, if the lease was not granted because the storage location is being leased by another client (step 10012), an exception handler is invoked that waits a predetermined amount of time (step 10014) and processing continues to step 10002.

An alternative, however, is that the server may queue lease requests. In this case, after waiting a predetermined amount of time, the client determines if it received a response indicating that the lease has been granted. If a lease is subsequently granted, processing continues to step 10024 in figure 10B. If a lease is not granted, the client waits and then continues to step 10002.

If the client determines the lease request was successful in step 10006, the client has an active lease. At this point, the client may access the storage locations covered by the lease (step 10024). After accessing the storage locations, the client determines if it is finished accessing the storage locations (step 10026). If the client is finished accessing the storage locations, the client determines if the lease expired (step 10028). If the lease expired, processing ends and no communication is necessary between the client and the server (i.e., natural termination occurred). Otherwise, if the lease is still active, the client invokes the cancel method (step 10030). The client does this for optimization purposes. The client accesses the cancel method via the lease object. The cancel method informs the server that the client is no longer interested in the storage locations. Accordingly, the cancel method allows the server to reclaim the storage location for use by other programs in an expeditious fashion.

If the client's use of the leased file is not completed, the client determines if the lease is about to expire (step 10032). This is achieved by the client comparing the current time with the duration of the lease. The duration of the lease is found by invoking the getDuration method.

If the lease is not about to expire, the client may continue to access the storage location in step 10024. However, if the lease is about to expire, the client must decide whether or not to renew the lease (step 10034). If the client chooses to renew the lease, the client invokes the renew method of the lease object. If the renew method is invoked, processing continues to step 10024. If the client does not renew the lease, the processing ends and no communication is necessary between the client and the server (i.e., natural termination).

Figure 11 depicts a flow chart of the steps performed by the server when a client requests a lease. These steps may be invoked when a client creates a file, requests a lease on a file that already has one, or invokes the renew method. The first step performed by the server is to receive a lease request by the client (step 11002). After receiving the request, the server examines the parameters to verify the propriety of the request (step 11004).

After examining the parameters, the server determines if the request is proper (step 11006). For example, the server checks if the requested storage location is, in fact, an actual storage location and if the client possesses a sufficient privilege level. Also, the server verifies that a desired lease period is specified. Additionally, the server checks if the type of access requested is available. If the server determines that the lease request is improper, the server generates an exception and invokes the appropriate client event handler (step 11008) and processing ends.

If the request is proper, the server executes the lease period algorithm ("LPA") to determine the lease period that should be granted (step 11010). The LPA produces a lease period that can range from the client's desired lease period to no lease at all. The LPA considers many factors in determining a lease period, these factors include, whether the request initiated from a renew method, a create instruction or a subsequent lease request, the client's usage patterns, the demand of the storage location, and the size of the storage locations, (if access to a large grouping is requested). Additionally, the LPA may consider the value of the storage locations. For example, if a particular storage location is very expensive to access or in high demand, the LPA may grant only short leases.

Once the LPA determines the lease period, the server determines if a lease period was granted (i.e., greater than zero) (step 11012). If no lease period was granted, the server generates an exception (step 11008) and processing ends. As described above, an alternative is that the

server may queue lease requests. In this case, the server stores the lease requests for later processing.

If the LPA did grant a lease period, the server determines if an exact lease was requested by the client (step 11016). If an exact lease was requested by the client, the server determines if the lease period granted by the LPA is less than the requested lease period (step 11018). If the lease granted by the LPA is less than the exact requested lease period, the server generates an exception (step 11008) and processing ends.

If an exact lease was not requested or an exact lease was granted, the server creates a lease object and returns it to the client (step 11020). It should be noted that the server, by monitoring active lease periods, reclaims leased storage locations that no longer have any active leases pertaining to them.

The foregoing description of an implementation of the invention has been presented for purposes of illustration and description. It is not exhaustive and does not limit the invention to the precise from disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practicing of the invention. For example, the described implementation includes software but the present invention may be implemented as a combination of hardware and software or in hardware alone. The scope of the invention is defined by the claims and their equivalents.

WHAT IS CLAIMED IS:

1. A method in a computer system having storage, comprising the steps of:

receiving an access request from a caller specifying a portion of the storage and specifying a requested lease period;

determining a lease period during which the caller has access to the portion of the storage;

advising the caller of the determined lease period; and

permitting the caller to access the portion of the storage for the determined lease period.

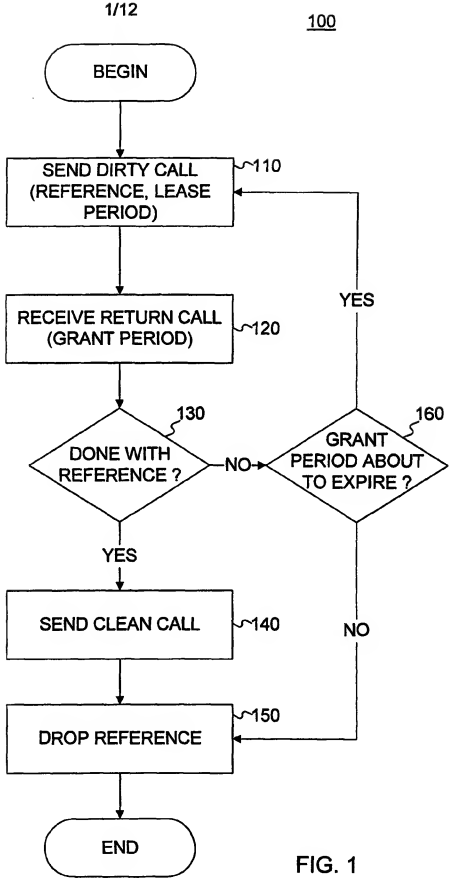


FIG. 1

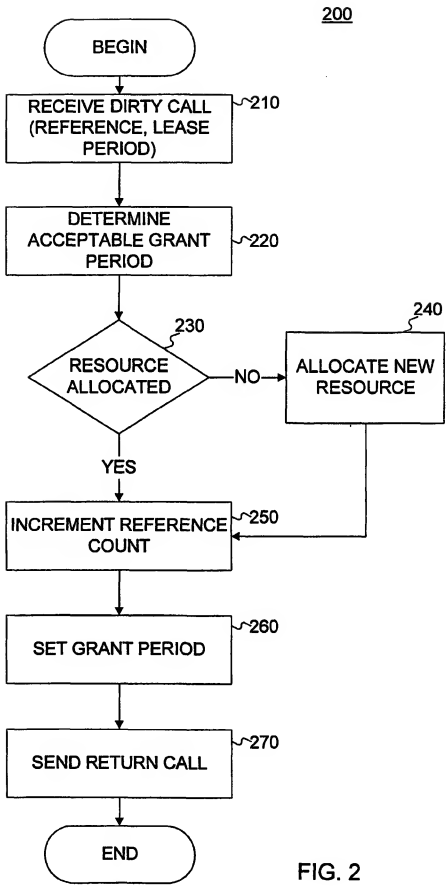


FIG. 2

300

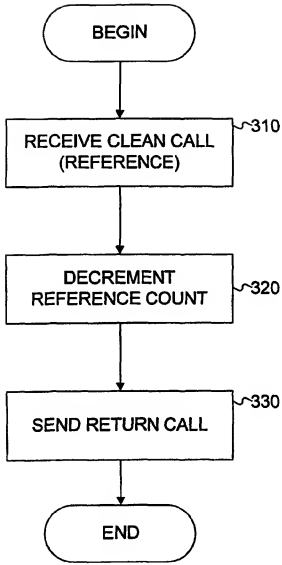


FIG. 3

400

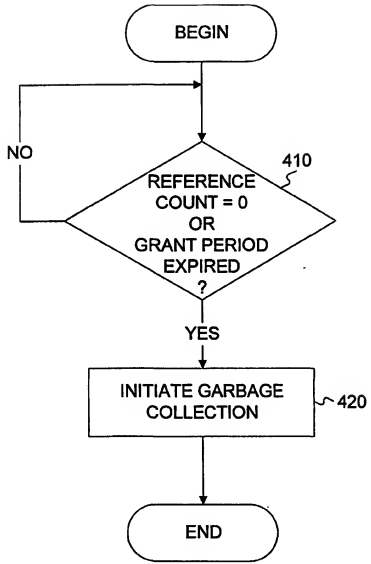
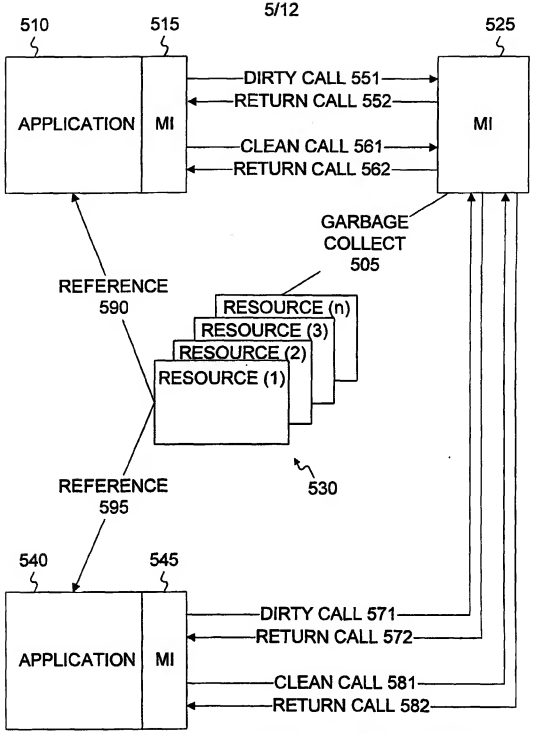


FIG. 4



—DIRTY CALL→	INCREMENT REFERENCE COUNTER
—CLEAN CALL→	DECREMENT REFERENCE COUNTER
REFERENCE COUNTER = 0	INITIATE GARBAGE COLLECTOR
GRANT PERIOD EXPIRED	INITIATE GARBAGE COLLECTOR

FIG. 5

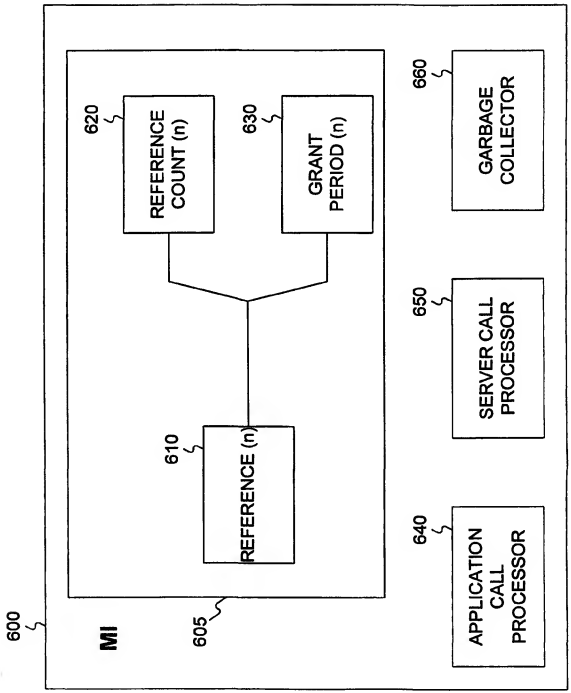


FIG. 6

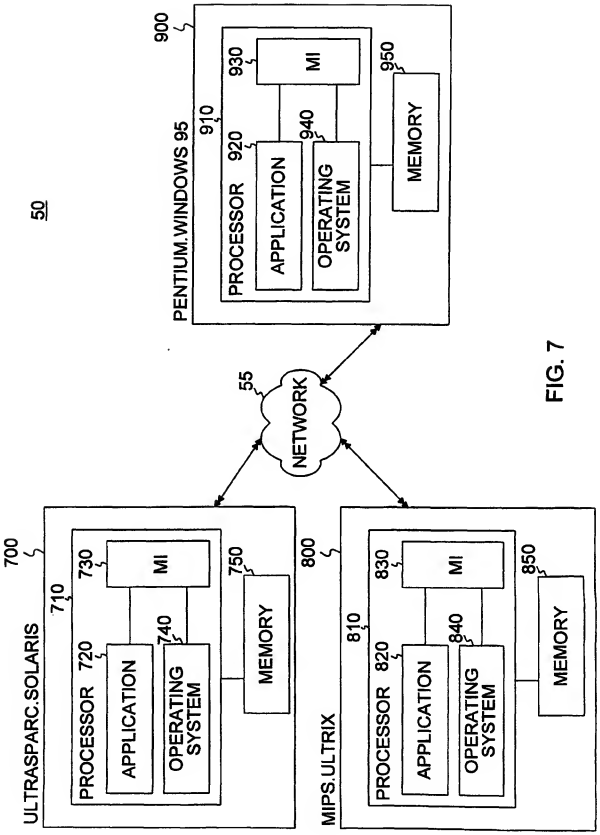


FIG. 7

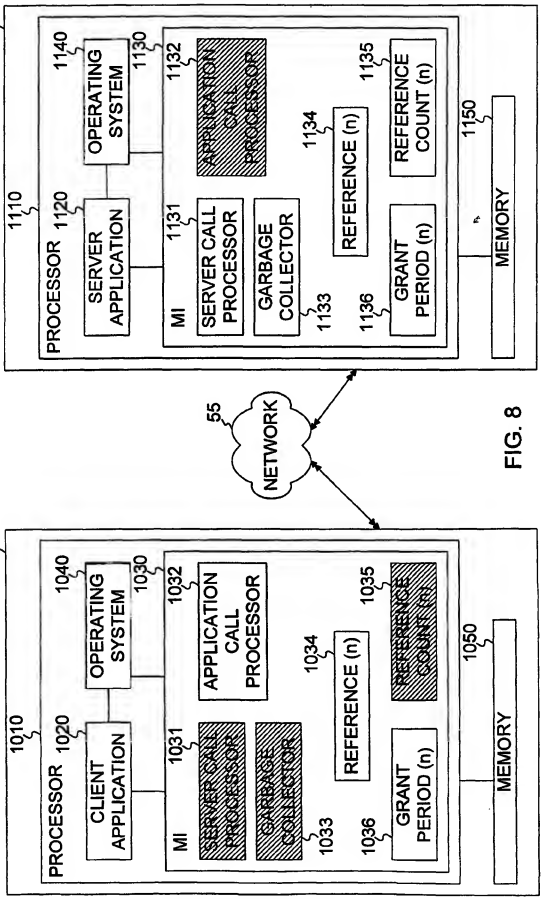


FIG. 8

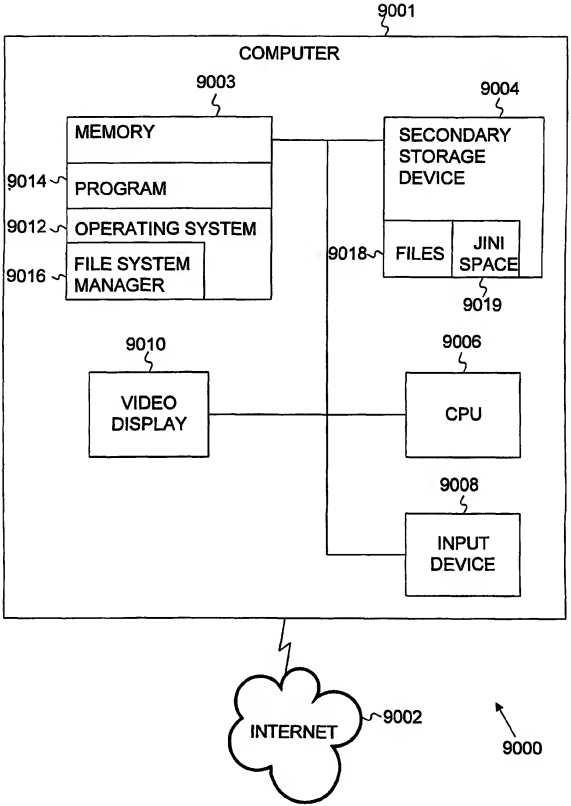


FIG. 9

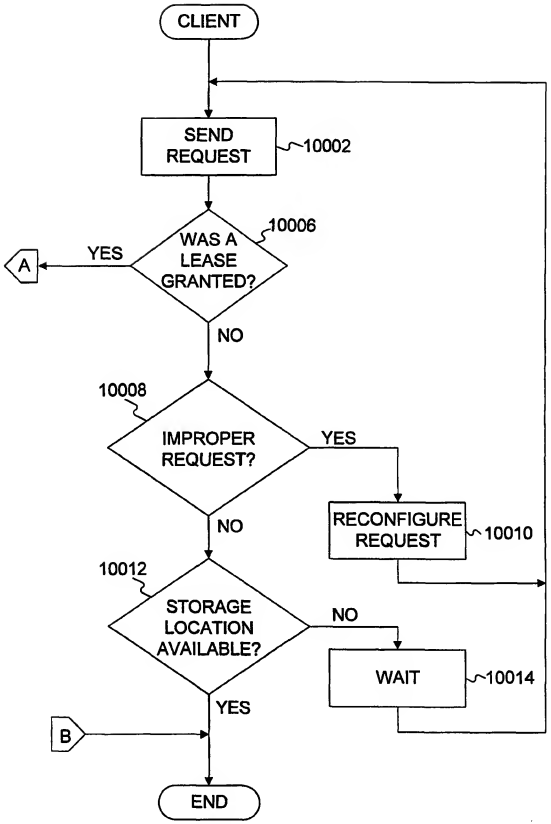


FIG. 10A

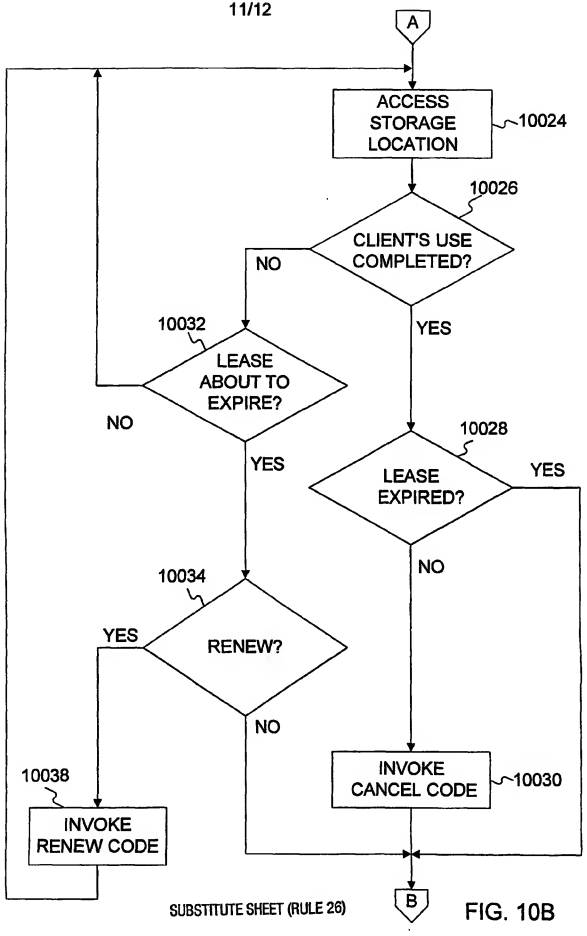


FIG. 10B

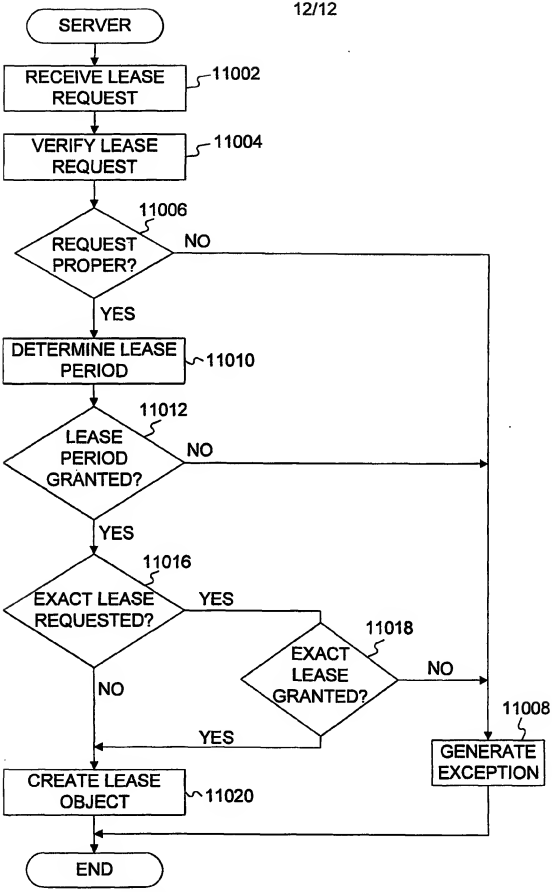


FIG. 11

INTERNATIONAL SEARCH REPORT

In. . atic Application No
PCT/US 99/03394

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F9/46

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 4 939 638 A (STEPHENSON R ASHLEY ET AL) 3 July 1990 (1990-07-03) column 1, line 35 - line 68	1
A	EP 0 569 195 A (ROCKWELL INTERNATIONAL CORP) 10 November 1993 (1993-11-10) column 3, line 21 - column 4, line 1	1
A	ANONYMOUS: "Resource Preemption for Priority Scheduling. November 1973." IBM TECHNICAL DISCLOSURE BULLETIN, vol. 16, no. 6, page 1931 XP002109435 New York, US the whole document	1

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
16 July 1999	29/07/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Brandt, J

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

Int. Appl. No. PCT/US 99/03394

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 4939638 A	03-07-1990	NONE	
EP 0569195 A	10-11-1993	US 5353343 A	04-10-1994
		CA 2095191 A	31-10-1993
		JP 6284458 A	07-10-1994